ACTA PHYSICA POLONICA A No. 1 Vol. **138** (2020)

Proceedings of the 10th Polish Symposium on Physics in Economy and Social Sciences (FENS 2019)

Measuring Strength of Detected Correlations between Incomes of Spouses

P. Łukasiewicz*, K. Karpio and A.J. Orłowski

Institute of Information Technology, WULS-SGGW,
Nowoursynowska 159, PL 02776 Warsaw, Poland

Doi: 10.12693/APhysPolA.138.101

*e-mail: piotr_lukasiewicz@sggw.edu.pl

This paper presents the studies of possible correlations between the incomes of women and men in the USA. In our previous paper, when we proposed a two-parameter model of spouses income distribution, we observed that the separate incomes of women and men could be correlated. In this paper we prove that such correlations do exist, by computing the standard Pearson correlation coefficient.

topics: personal and family income distributions, correlations between incomes of spouses

1. Introduction

In paper [1] we studied interdependences between income distribution for families with two adults atributions of personal incomes of males and females. We used microdata for years from 2001 to 2016 collected within the project Current Population Survey (CPS) in USA [2]. We considered the main part of the income distribution for families, covering about 98%–99% of objects depending on year. We omitted a tail of the distribution (1%–2%), which had the other shape than the main part. Now, we assumed that the personal income distributions of males and females have exponential shapes. They can be approximated by the density function

$$P(x;a) = \frac{\exp(-x/a)}{a} \quad x > 0, \quad a = \langle x \rangle,$$

where x is an income (see [3, 4] and Fig. 1).

We also proposed in [1, 5] two-parametric model of income distribution for families with two adults defined as the convolution of two exponential distributions:

$$P_2\left(x;a,b\right) = \frac{\exp\left(-x/a\right) - \exp\left(-x/b\right)}{\left(a-b\right)}, \qquad (1)$$

where x is positive income and a, $b > 0$ are parameters of exponential distributions. In fact, the formula derived in [1] is incorrect, it should have a form of (1), see [5]. Let's note, when $X \sim P_2$, then $\mathrm{Var}\left(X\right) = a^2 + b^2$. Thus, the variance is equal to the sum of the variances of exponential distributions. If a and b are parameters of personal income distributions of males and females, then model (1) will describe distribution of 92%–95% of the population well, depending on year. On the other hand, if these model parameters are evaluated by extrapolating the function (1) to the income distribution for families, then the model will explain about 98%–99% of the population. The results for years 2006 and 2015 are presented in Fig. 2. We showed that the differences are relatively small during the years 2001–2009, and significantly rising after 2009. We believe that the observed differences indicate correlation between the male and female incomes. In this paper we study the correlations quantitatively, calculating the Pearson correlation coefficient.

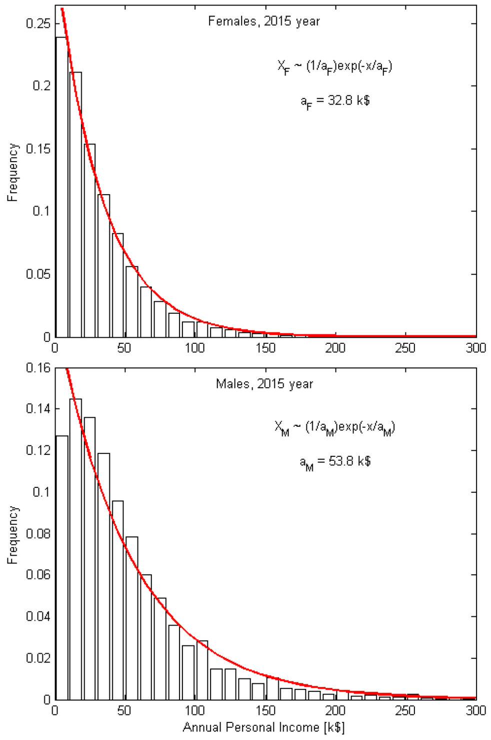

Fig. 1. Empirical personal income distributions and exponential fits for females and males in 2015.

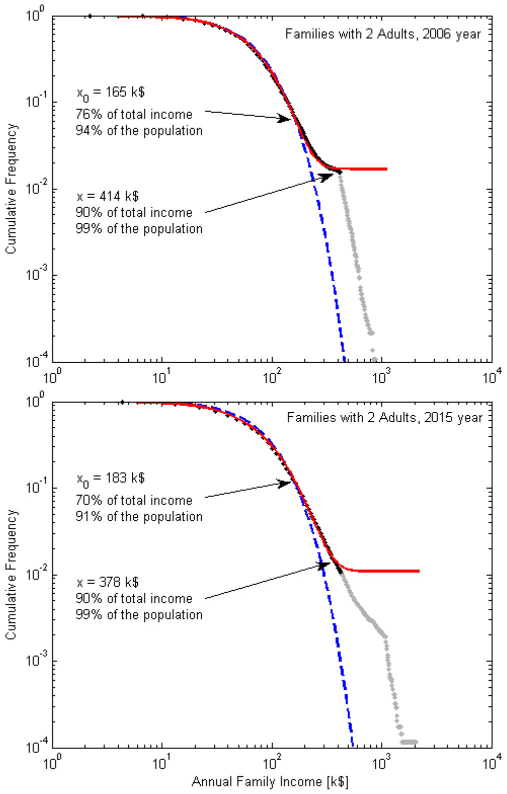

Fig. 2. Cumulative income distributions of families with 2 adults in 2006 and 2015 with models (1) in log-log scale. The dashed line represents the convolution of two exponential functions (for men and women). The solid line is a result of the model (1) fitted to the data. The grey data point were excluded from data.

2. Correlations between incomes of spouses

To estimate the Pearson correlation coefficient we use the following theorem. For any random variables X and Y:

$$\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)$$

$$+2\mathrm{Cov}\,(X,Y)\,, \tag{2}$$

where $\mathrm{Cov}\,(X,Y)$ indicates the covariance of the random variables X and Y.

Let the random variables X_M and X_F represent the income of male and female, respectively. We assume they both have exponential shapes: $X_M \sim a_M^{-1}\exp(-x/a_M)$ and $X_F \sim a_F^{-1}\exp(-x/a_F)$. Based on (2) one can write: $\mathrm{Var}\,(X_M + X_F) = Var\,(X_M) + Var\,(X_F) + 2\mathrm{Cov}\,(X_M, X_F)$, or simpler as:

$$\mathrm{Var}\,(X_M + X_F) = a_M^2 + a_F^2$$

$$+2\mathrm{Cov}\,(X_M, X_F)\,. \tag{3}$$

Note that $a_M^2 + a_F^2$ is equal to a variance of the convolution $P_2(x; a_M, a_F)$. Since this function does not describe the whole income distribution of families, then $\mathrm{Cov}\,(X_M, X_F) \neq 0$.

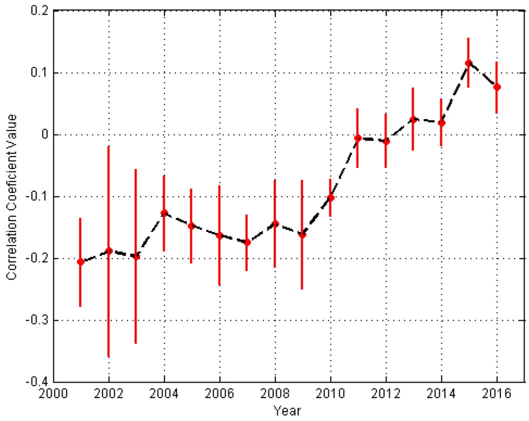

Fig. 3. Values of the correlation coefficients vs. year. Vertical lines indicate 95% confidence intervals.

On the other hand, when neglecting in model (1) the minor discrepancies between the model and data at the right end of the distribution, we obtain: $(X_M + X_F) \sim P_2\,(x; \alpha, a_1, a_2)$, where a_1, a_2 are evaluated parameters and α is a normalizing parameter. Next, using the formula $\mathrm{Var}\,(X) = E\left(X^2\right) - \left(E\,(X)\right)^2$ it is easy to show that:

$$\mathrm{Var}\,(X_M + X_F) = \alpha\,(2-\alpha)\left(a_1^2 + a_2^2\right)$$

$$+\alpha\,(2-2\alpha)\,a_1 a_2 \tag{4}$$

We obtain a formula for covariance $\mathrm{Cov}\,(X_M, X_F)$ from (3) and (4):

$$\mathrm{Cov}\,(X_M, X_F) = \frac{\alpha(2-\alpha)}{2}\left(a_1^2 + a_2^2 + a_1 a_2\right)$$

$$-\frac{1}{2}\left(\alpha^2 a_1 a_2 + a_M^2 + a_F^2\right). \tag{5}$$

Dividing both sides of this formula by standard deviations of exponential distributions we obtain:

$$\mathrm{Corr}\,(X_M, X_F) = \frac{\alpha(2-\alpha)}{2a_M a_F}\left(a_1^2 + a_2^2 + a_1 a_2\right)$$

$$-\frac{1}{2a_M a_F}\left(\alpha^2 a_1 a_2 + a_M^2 + a_F^2\right). \tag{6}$$

Using (6) we calculate values of the correlation coefficients between incomes of males and females for years from 2001 to 2016. Errors are estimated using the formula for mean square deviation. The results are listed in Table I and in Fig. 3, where we present dependence of the values of the correlation coefficient on time. The vertical lines on the plot represent doubled errors which are the approximation of the 95% confidence intervals. The big values of the errors for 2002 and 2003 are caused by relatively big errors of fitted parameters of the model (1).

We observe values of the correlation coefficient from about −0.2 to 0.1. The values are negative till 2010, statistically consistent with zero between 2011 and 2014, and positive for the last two years. The value of the correlation seems to increase after 2009 going from negative to positive. The direction of the relation between incomes of spouses changed after 2014. Negative correlations mean

TABLE I

Estimated values of the correlation coefficients between incomes of males and females in USA.

Year	Corr. coefficient	Error
2001	−0.207	0.035
2002	−0.189	0.084
2003	−0.198	0.070
2004	−0.128	0.030
2005	−0.148	0.030
2006	−0.164	0.040
2007	−0.175	0.023
2008	−0.145	0.035
2009	−0.162	0.044
2010	−0.103	0.015
2011	−0.006	0.024
2012	−0.011	0.021
2013	+0.024	0.025
2014	+0.019	0.019
2015	+0.116	0.019
2016	+0.076	0.021

the difference between incomes of spouses are relatively big (low and high incomes come together). Similarly, positive correlations mean the incomes of spouses are closer to each other. Let note, that obtained results show an existence of the relatively week correlations, predominantly between −0.1 and 0.1. Nevertheless, the results indicate some sociological changes in American society after 2009. At this stage of the studies it is difficult to point out socio-economic reasons of such behavior, it requires further investigation. It appears, that it is not possible at the statistical level of the analysis of distributions. We do not address this issue in this work, it requires data about personal incomes of the family members.

However, we can provide an statistical explanation of the observed changes of the correlation coefficient's sign. According to (5) sign of the correlation coefficient is a difference of a variance of the distribution $P_2\left(x;\alpha,a_1,a_2\right)$ and a sum of variances of the exponential distributions: $a_M^2+a_F^2$. We concluded in [1] that for family incomes below certain x_0, incomes of spouses are independent and correlations arise beyond the threshold x_0 (see Fig. 2, data point, at which the curves diverge). That would be true if cumulative functions of the fitted model and the convolution have been in agreement with each other for incomes $x<x_0$. Besides that, one would observe only positive correlations because convolution drops faster thus having smaller variance (see Fig. 2). These are not consistent with our results, plots of cumulative distributions does not show differences in left and middle part of the distributions. It turns out that the fitted model and the convolution are not in agreement with each other below x_0 and observed differences are significant for the majority of years. We show the results graphically in Fig. 4, where theoretical density curves and histogram representing left and middle part of the empirical distribution are shown. The fitted model is not compatible with the convolution, while describing empirical distribution better. The curve representing convolution is moved right and is wider than the fitted model. A magnitude of the observed discrepancies is different for various years, but the model fitted to data is more concentrated than the convolution of the distributions for the majority of years. Density is bigger around the dominant point and is smaller on the slope. Conversely, the fitted model reaches further than the convolution thus having a bigger variance in the range the highest incomes. The direction and value of the correlation coefficient result from the opposite differences between models in low − middle and high range of incomes. Therefore, values of the correlation coefficient being evaluated with means of the theoretical distributions (in the range $0\div\infty$) are the consequences not only of the differences above income x_0 but also of differences below x_0. It seems that correlations between incomes of the family members occur in the ranges of low and mean incomes of family members, at least for some years. There are necessary more detailed studies, which will allow to find out the income threshold for correlations occurrences.

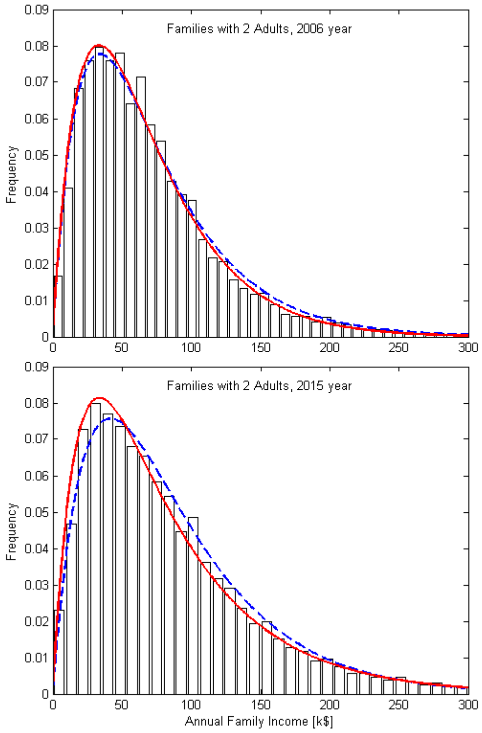

Fig. 4. Income distributions of families with 2 adults in 2006 and 2015 with models (1). The dashed line represents the convolution of two exponential functions (for men and women). The solid line is a result of the model (1) fitted to the data.

3. Summary

In this paper one elaborated the formula to evaluate value of the correlation coefficient between incomes of males and females in families with two adults. The formula was obtained based on two-parametric model of income distribution for families. Using the survey microdata for years 2001–2016 in USA one showed that the correlations are going from negative to positive, being relatively small. We want to prove the results utilizing simulation methods. We are going to evaluate the correlations independently by generating the components of the family incomes We also want to evaluate a limit income of the family, above which correlations occur. The further studies will be conducted in this direction.

References

[1] P. Łukasiewicz, K. Karpio, A.J. Orłowski, *Acta Phys. Pol. A* **133**, 1441 (2018).

[2] The USA Census Data.

[3] A.A. Drăgulescu, V.M. Yakovenko, *Eur. Phys. J. B* **20**, 585 (2001).

[4] P. Łukasiewicz, K. Karpio, A.J. Orłowski, *Acta Phys. Pol. A* **121**, B-82 (2012).

[5] P. Łukasiewicz, K. Karpio, A.J. Orłowski, *Acta Phys. Pol. A* **133**, 1441 (2018), ERRATUM: *Acta Phys. Pol. A* **137**, 436 (2020).